YData Seminar: Statistics and Data Science 171

An Introduction to Text Data Science

Thursday, January 17

Outline

- Overview of course
- Syllabus and logistics
- Demo and Lab

Context

- Text is everywhere
- Increasingly available in electronic form
- It's the "diary" of human activity

Traditional data

- Tabular data
- Time series
- Surveys
- Recordings of experiments
- Economic indicators
- *Text is different*

Text as Data

Matthew Gentzkow
Stanford

Bryan T. Kelly
Chicago Booth

Matt Taddy
Chicago Booth and Microsoft Research

Abstract

An ever increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications.

This course

- An introduction to text analysis
- Not (really) NLP
- Some “elementary” (direct) methods
- Some “advanced” (ML) methods

Goals

- Give you confidence and “comfort level” with text
- Starting point for using text in your intellectual pursuits (senior project, research, startup, side projects...)
- Strenghthen programming skills
- Introduction to some advanced ideas (topic models, embeddings...)

Organization

- Coupled with YData
- Same computing platorm
- Each class meeting will be based on a Jupyter notebook (Lab)

Seminar Scope

- Half course credit
- QR requirement
- 20-25 students
- Please complete survey on Canvas
- Roster set by early next week
- No auditors (sorry!)

Staff

- JL
- Yi Chern Tan (ULA)
- We will hold weekly office hours (TBD)

Seminar structure

- Every meeting will be based on a Lab
- A lab is a Jupyter/Python notebook
- First 20-40 minutes: Discussion of background
- Next 20-30 minutes: Begin on lab
- Remaining time: Independent work on labs
- Complete lab by next class period
- Cautious estimate: $\sim$5 hours outside work

Additional assessment

- Midterm (in class March 7): Like a lab
- Project: Your own lab!
- Final (to be determined)

Distribution

- Labs: 50%
- Midterm: 15%
- Project: 20%
- Final: 15%

My hope

“fun, cool stuff, useful for my later work, challenging but manageable, sparked my interest, ...”

The data – 2 labs each

- Gutenberg books
- State of the Union Addresses
- Scientific Articles
- Wikipedia
- Product reviews

Mindset

- This is an experiment
- Computing structure will require some teaks
- Seminar scope will require some tweaks
- Needed: Patience, feedback, interaction...

Questions?

Remaining time

- Preview of Gutenberg labs
- Lab 1: Python expressions (from YData)

http://ydata123.org/sp19/tds.html